intel ai summit
英特爾AI科技論壇

Bringing AI Everywhere

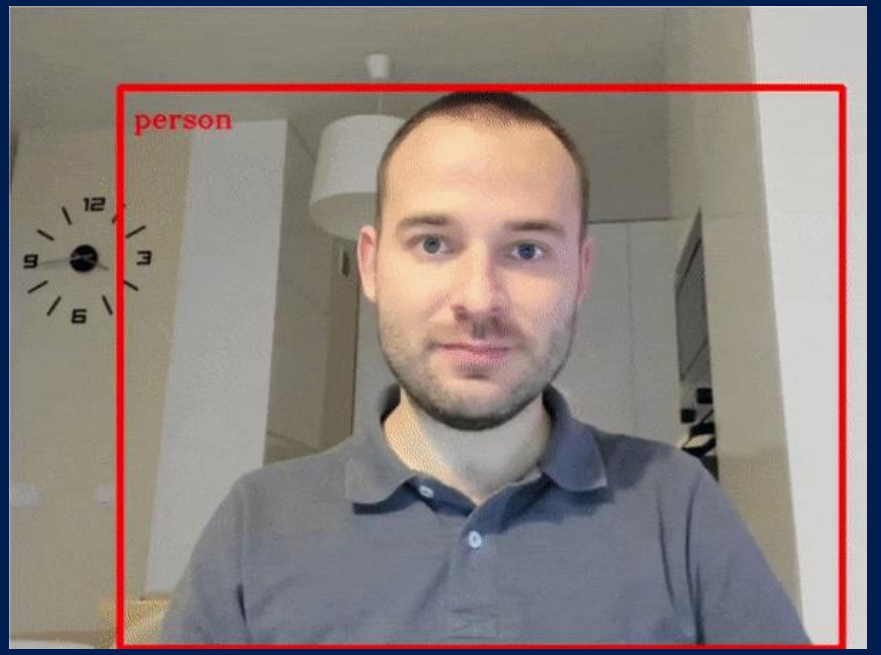
OpenVINO™ 2024.0 for Faster, Smarter Inference

Zhuo Wu
AI Software Evangelist

March 27th, 2024

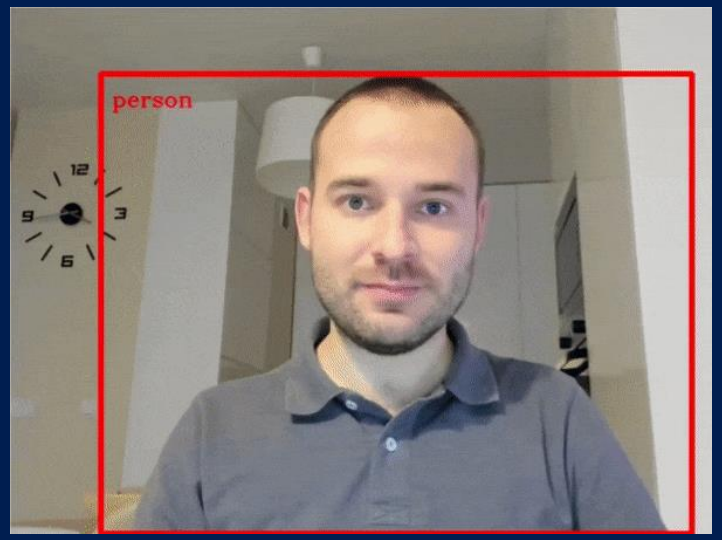
What can AI do?

Object Detection 🔍

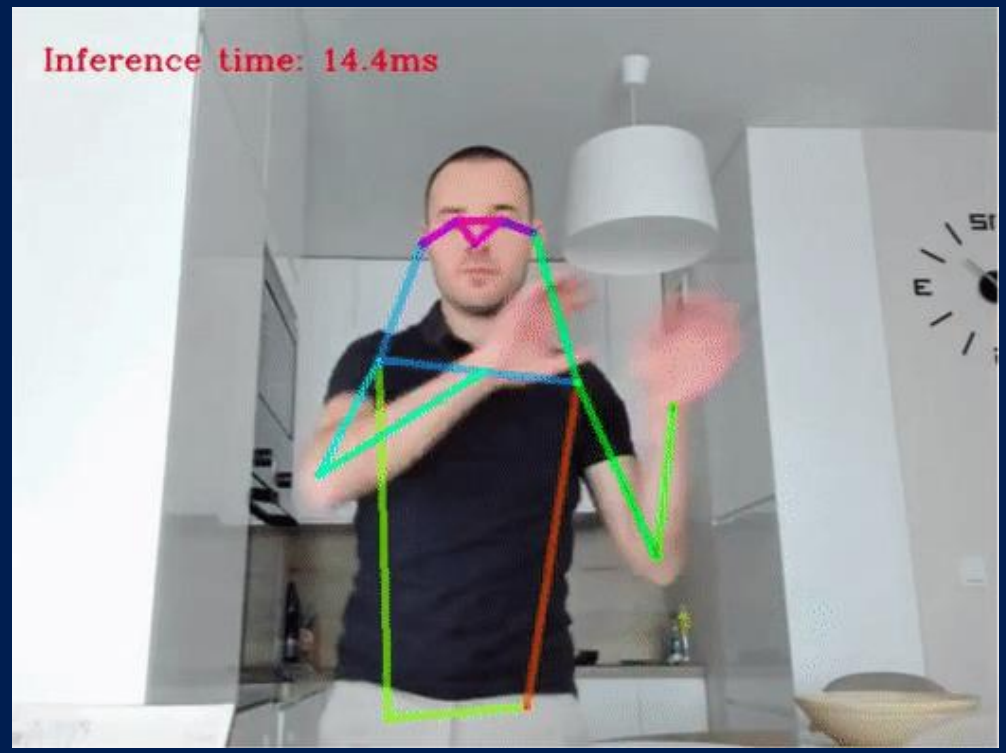


What can AI do?

Object Detection



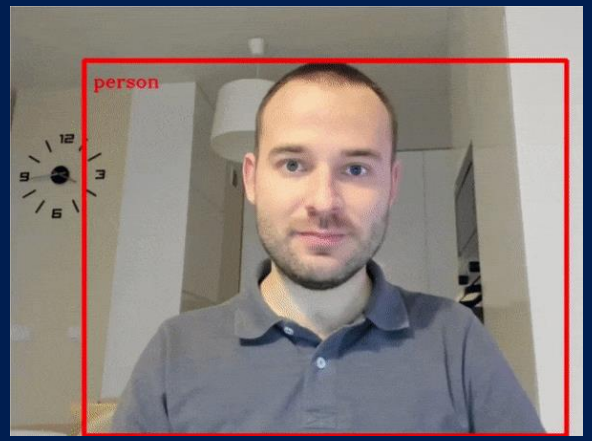
Pose Estimation



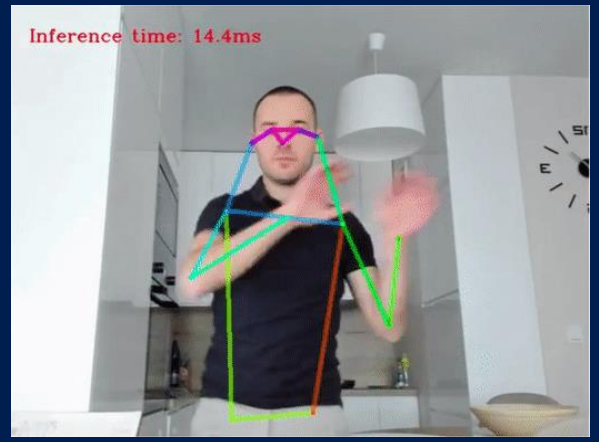
intel Ai summit

What can AI do?

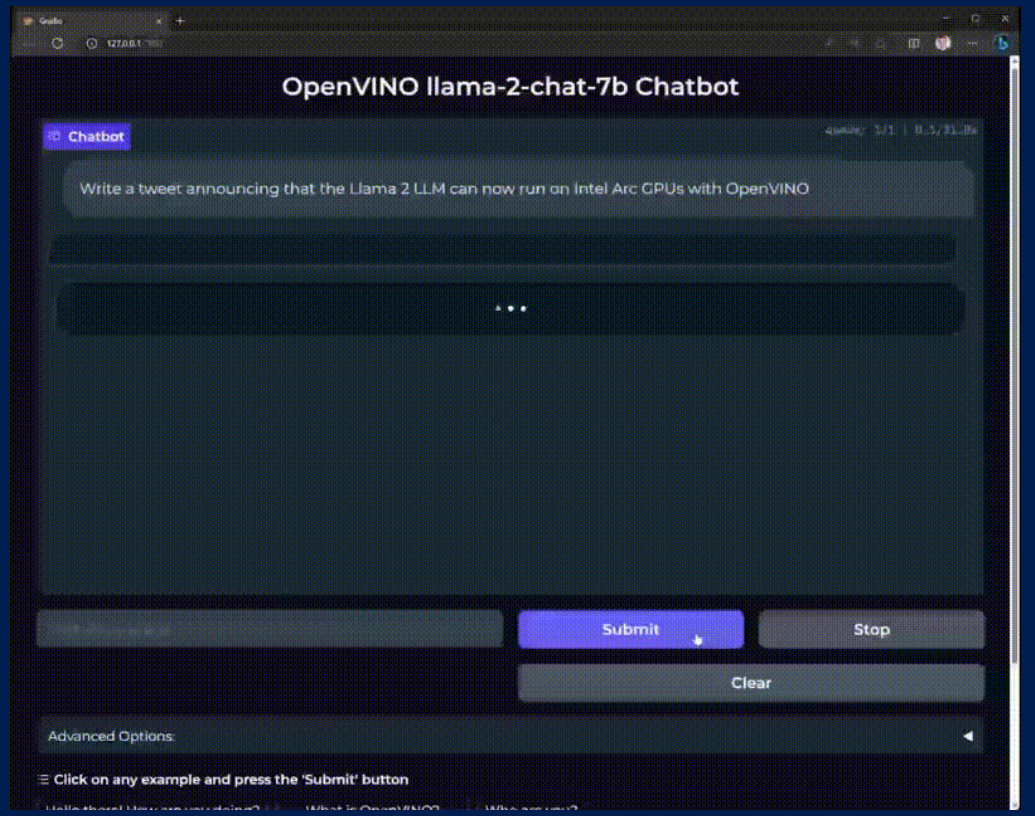
Object Detection



Pose Estimation



Llama2 Chatbot



intel Ai summit

The Challenge

Having Your Model Trained is Only 1/2 the Challenge!

Deployment is hard… and we know, because we've done a lot of it!

How Do I Deploy My AI Solutions?

Should we start with
CPU, GPU, or NPU?

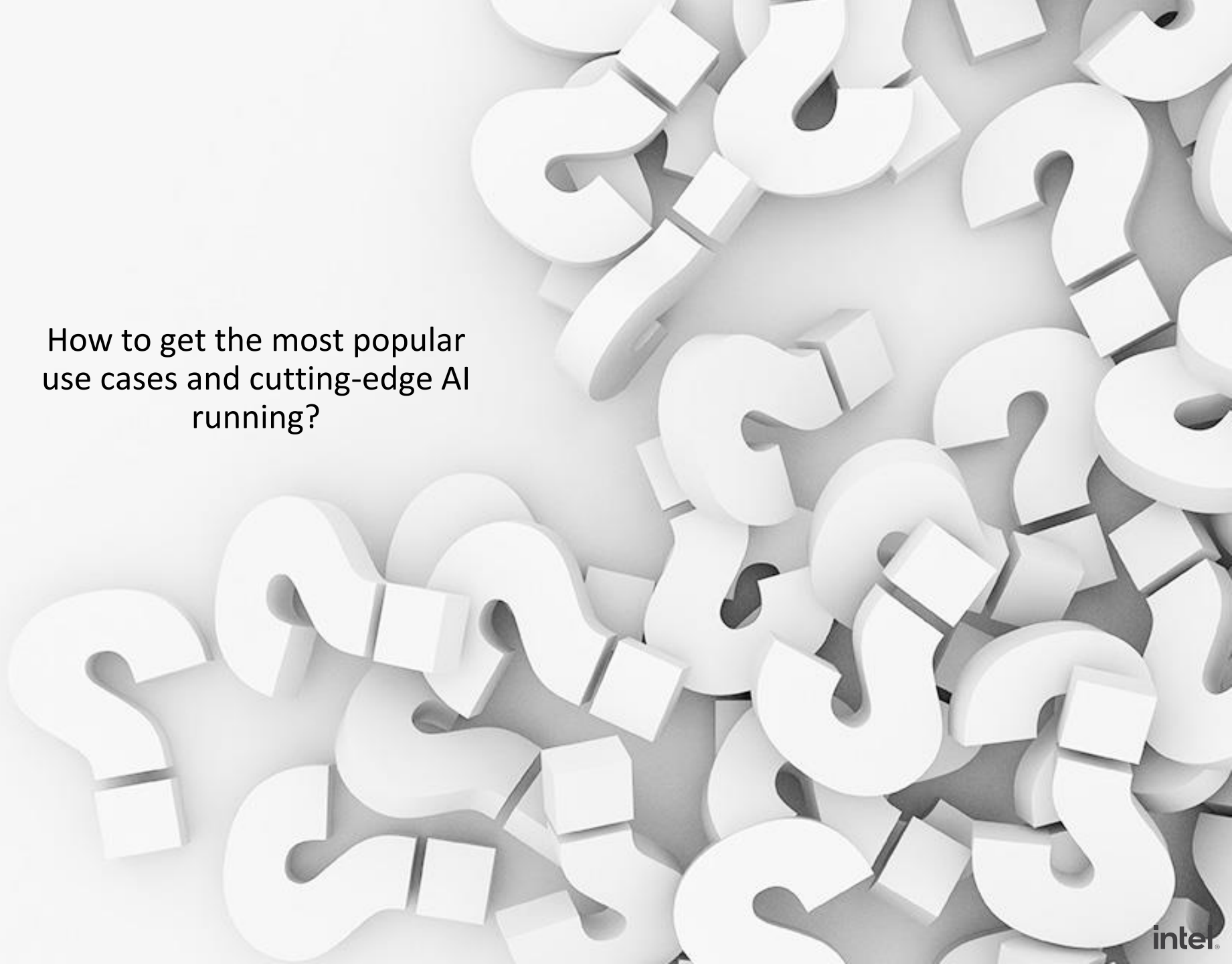
How about XPU or 'all of the above' since Intel now has CPU, GPU, and NPU all working under the same one roof.

OpenVINO™

How to get the best AI application performance out of the available hardware?

How to get the most popular use cases and cutting-edge AI running?

How to maintain code across different architectures?



OpenVINO
The Solution
intel

Google Summer of Code

open source initiative® Apache 2.0

OpenVINO™

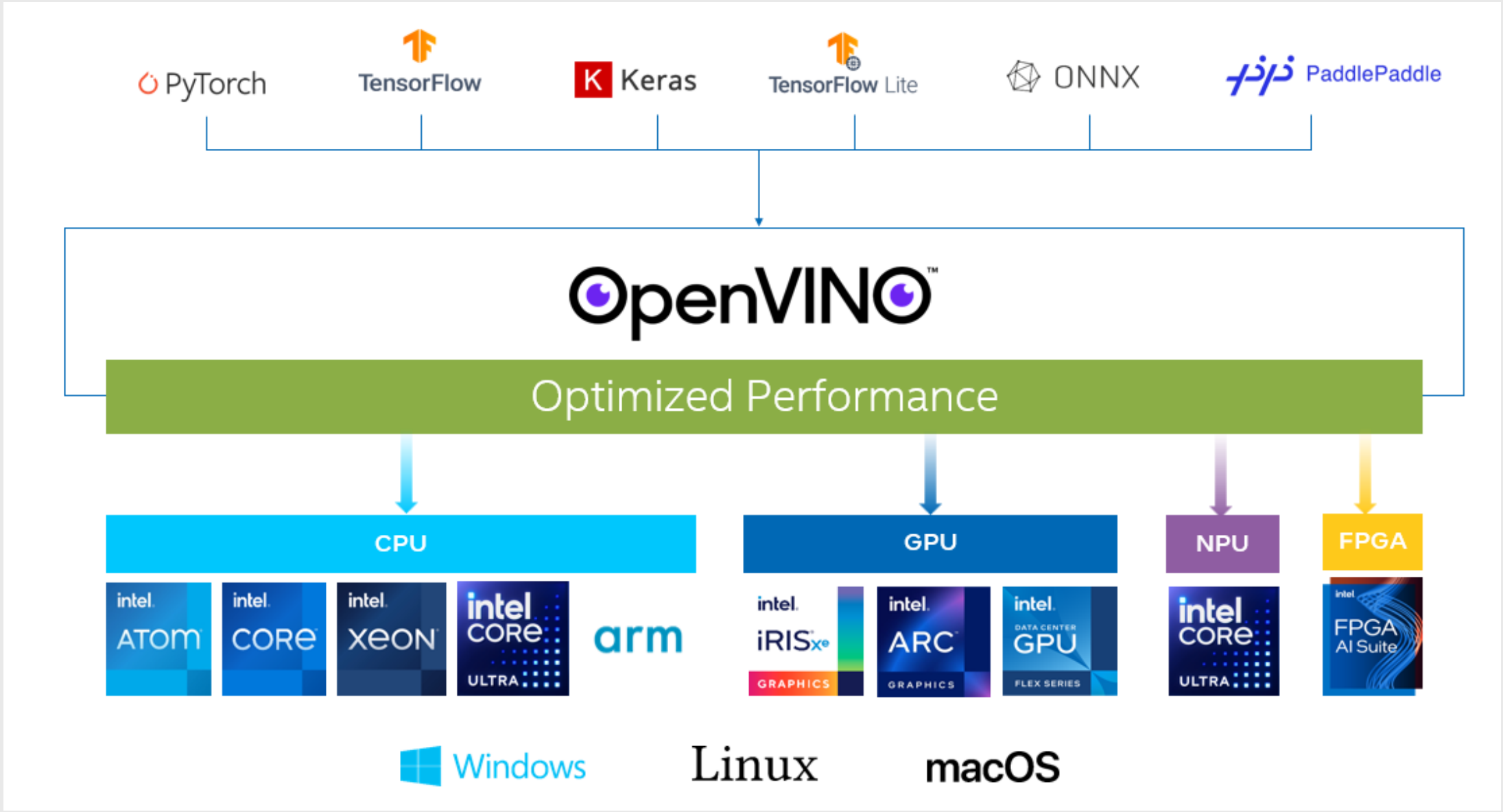
Visual
+
NLP, Audio, LLM, GenAI, Transformers...

and
Neural
Network

Optimization

Inference

Powered by oneAPI


PyTorch
TensorFlow
Keras
TensorFlow Lite
ONNX
PaddlePaddle
OpenVINO™
Optimized Performance
CPU
GPU
NPU
FPGA
intel ATOM
intel CORE
intel XEON
intel CORE ULTRA
arm
intel iRISXe GRAPHICS
intel ARC GRAPHICS
intel DATA CENTER GPU FLEX SERIES
intel CORE ULTRA
intel FPGA AI Suite
Windows
Linux
macOS

OpenVINO™ 2024.0

Making it easier to deploy and accelerate Gen AI & LLMs

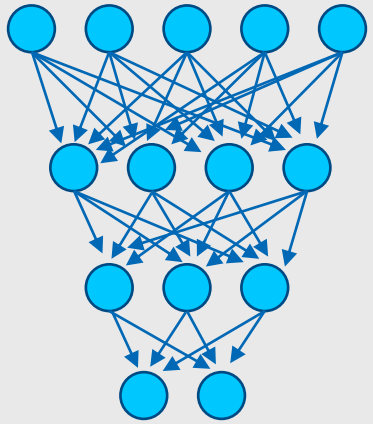
- **Improvement for LLMs**
- Simpler Workflow
- Enhancements for existing platforms
- More Deep Learning Models

intel Ai summit

Neural Network Compression Framework (NNCF)

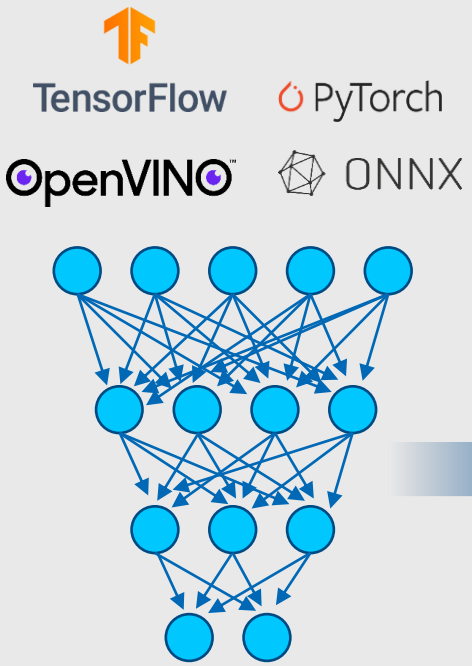
TensorFlow PyTorch

OpenVINO ONNX



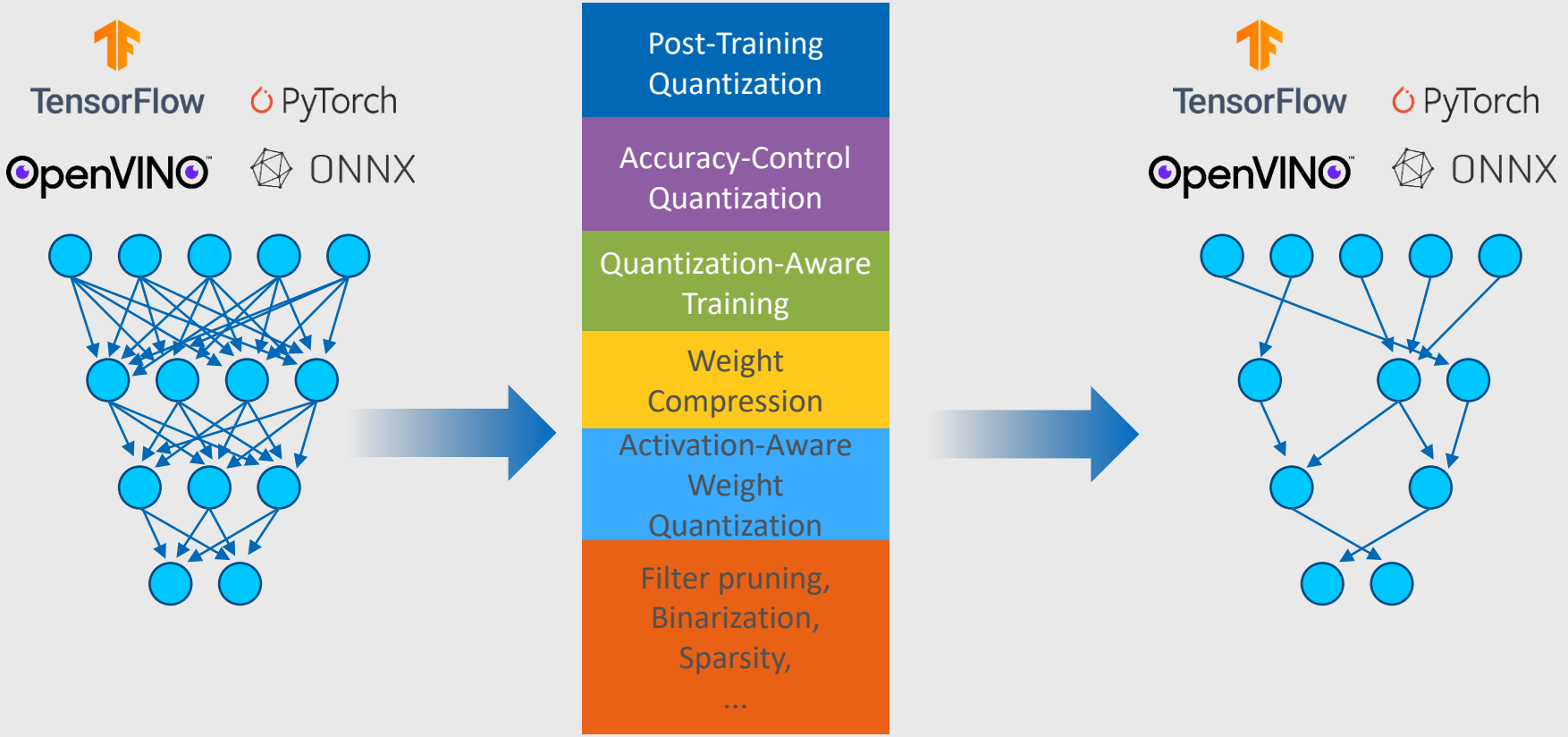
intel Ai summit

Neural Network Compression Framework (NNCF)



- Post-Training Quantization
- Accuracy-Control Quantization
- Quantization-Aware Training
- Weight Compression
- Activation-Aware Weight Quantization
- Filter pruning, Binarization, Sparsity, ...

Neural Network Compression Framework (NNCF)




```
pip install nncf
```

intel AI summit

Weight Compression for LLMs (NNCF)

Model	Mode	Perplexity	Perplexity Increase	Model Size (GB)
databricks/dolly-v2-3b	fp32	5.01	0	10.3
databricks/dolly-v2-3b	int8	5.07	0.05	2.6
databricks/dolly-v2-3b	int4_asym_g32_r50	5.28	0.26	2.2
databricks/dolly-v2-3b	nf4_g128_r60	5.19	0.18	1.9
meta-llama/Llama-2-7b-chat-hf	fp32	3.28	0	25.1
meta-llama/Llama-2-7b-chat-hf	int8	3.29	0.01	6.3
meta-llama/Llama-2-7b-chat-hf	int4_asym_g128_r80	3.41	0.14	4.0
meta-llama/Llama-2-7b-chat-hf	nf4_g128	3.41	0.13	3.5
togethercomputer/RedPajama-INCITE-7B-Instruct	fp32	4.15	0	25.6
togethercomputer/RedPajama-INCITE-7B-Instruct	int8	4.17	0.02	6.4
togethercomputer/RedPajama-INCITE-7B-Instruct	nf4_ov_g32_r60	4.28	0.13	5.1
togethercomputer/RedPajama-INCITE-7B-Instruct	int4_asym_g128	4.17	0.02	3.6

Significant Reduction in RAM usage!

intel Ai summit

Weight Compression for LLMs (NNCF)

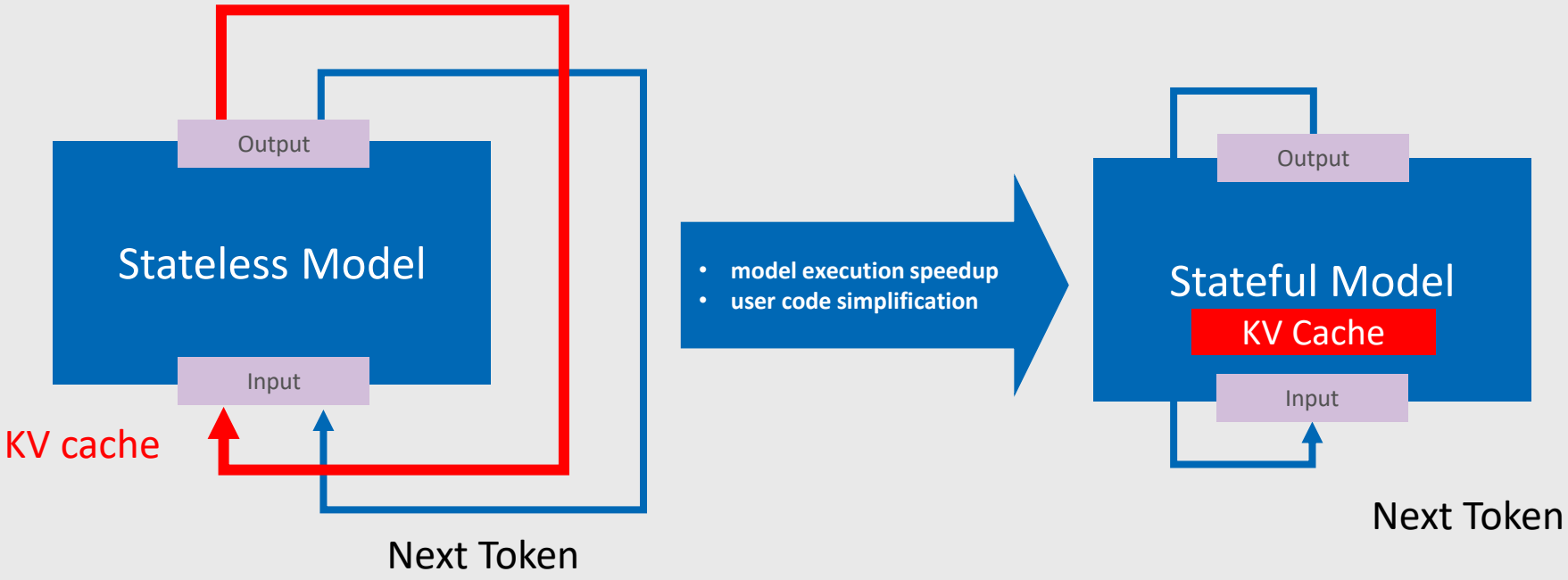
Model	100% int4, symmetric mode	lambada_openai Perplexity (↓)	lambada_openai Accuracy (↑)	Gsm8k Exact Match, 5 shot (↑)
stablelm-2-zephyr-1_6b	awq	**7.905**	**54.9**	**34.12**
stablelm-2-zephyr-1_6b	data-free	8.584	54.28	29.95
stable-zephyr-3b-dpo	awq	**8.4099**	**57.27**	**48.6**
stable-zephyr-3b-dpo	data-free	9.3011	56.18	46.17
zephyr-7b-beta	awq	**3.4564**	**71.05**	-
zephyr-7b-beta	data-free	3.5021	70.7	-

Better Quality of the Compressed Model!

intel Ai summit

Stateful Transformation

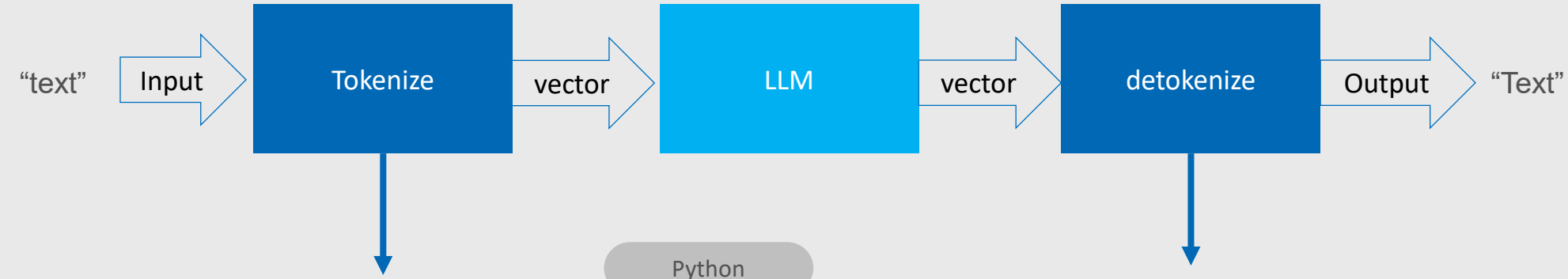
How to optimize KV cache with OpenVINO™



intel Ai summit

OpenVINO™ Tokenizers

3rd party dependencies **needed to use models with text inputs, e.g., LLMs, Stable Diffusion**



Python

```
from transformers import AutoTokenizer

tokenizer = AutoTokenizer.from_pretrained(model_id)

input_text  = tokenizer(prompt, return_tensors="np")
```

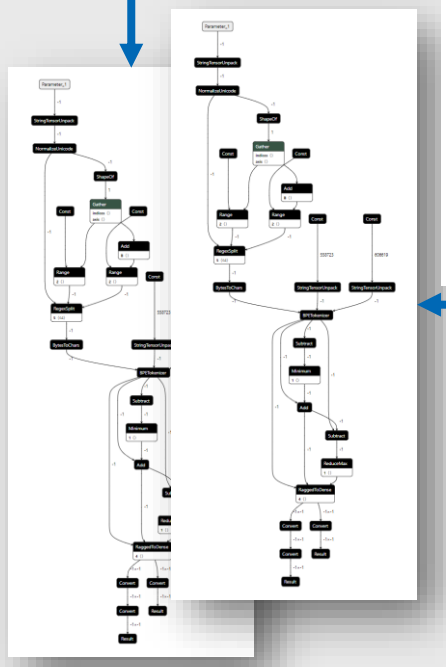
Python

```
output_text = tokenizer.decode(output_ids, skip_special_tokens=True)
```

intel Ai summit

OpenVINO™ Tokenizers

OpenVINO Tokenizers is a binary extension with text-processing operations and a python package that helps to obtain **tokenizer models**.



Huggingface Tokenizer Type	Tokenizer Model Type	Tokenizer	Detokenizer
Fast	WordPiece	✅	❌
	BPE	✅	✅
	Unigram	❌	❌
Legacy	SentencePiece .model	✅	✅
Custom	tiktoken	✅	✅

https://github.com/openvinotoolkit/openvino_contrib/tree/master/modules/custom_operations/user_ie_extensions/tokenizer/python

intel Ai summit

OpenVINO mistral-7b Chatbot

Chatbot

Chat Message Box

Submit

Stop

Clear

Advanced Options:

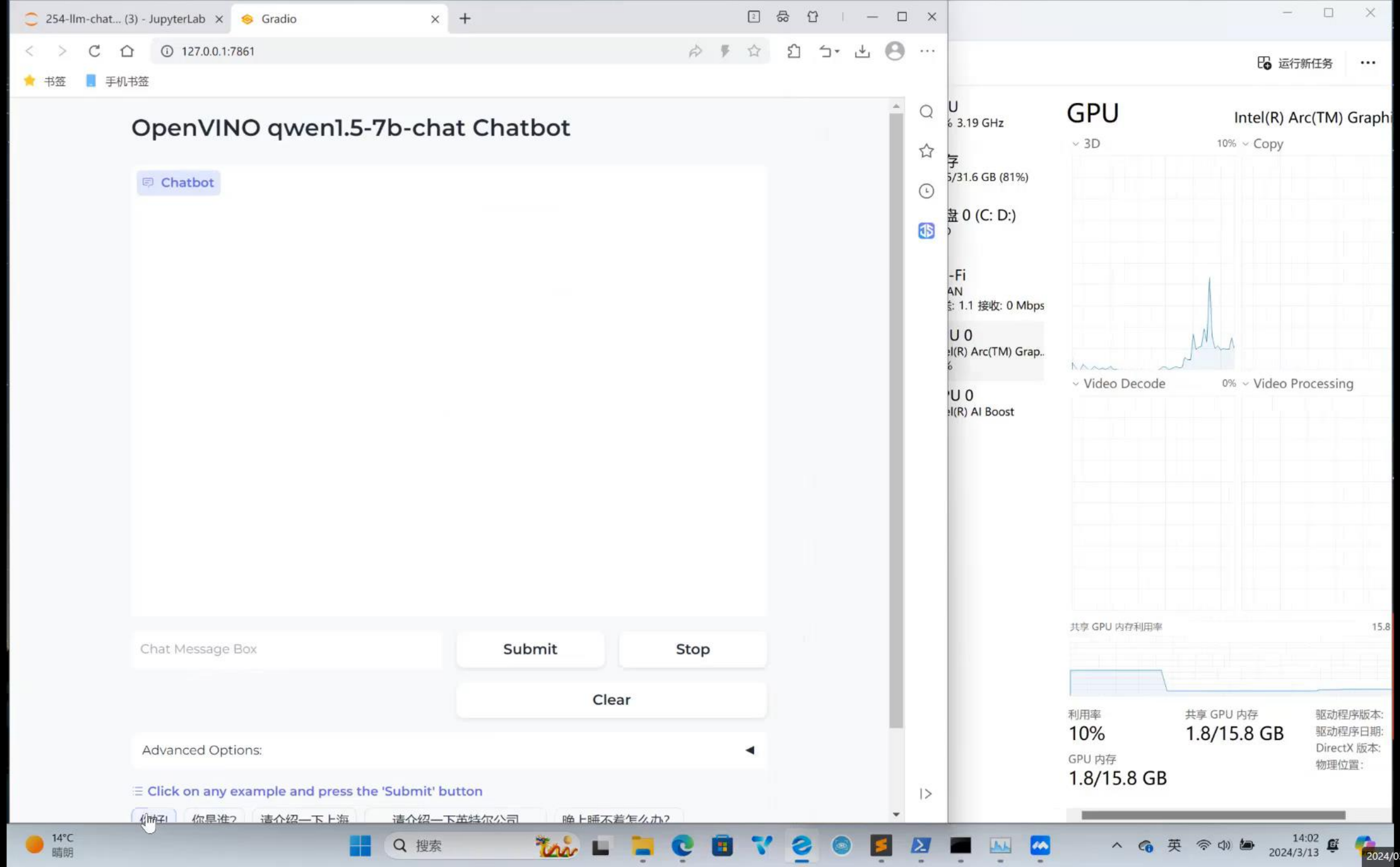
Click on any example and press the 'Submit' button

Hello there! How are you doing?

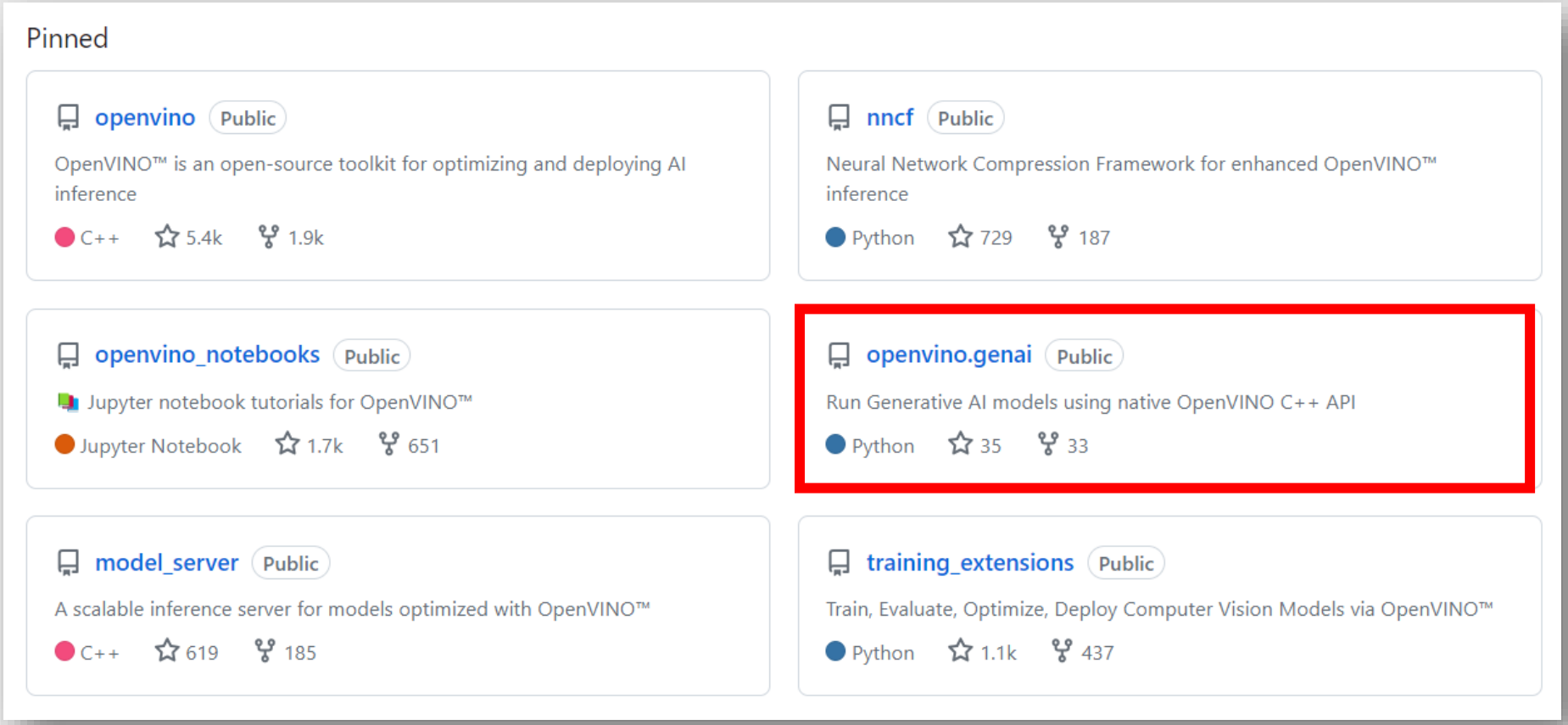
What is OpenVINO?

Who are you?

Can you explain to me briefly what is Python programming language?


254-llm-chat... (3) - JupyterLab
Gradio
127.0.0.1:7861
书签
手机书签
OpenVINO qwen1.5-7b-chat Chatbot
Chatbot
Chat Message Box
Submit
Stop
Clear
Advanced Options:
Click on any example and press the 'Submit' button
你好!
你是谁?
请介绍一下上海
请介绍一下英特尔公司
晚上睡不着怎么办?
运行新任务
GPU
Intel(R) Arc(TM) Graphi
3D
10%
Copy
Video Decode
0%
Video Processing
共享 GPU 内存利用率
15.8
利用率
10%
共享 GPU 内存
1.8/15.8 GB
GPU 内存
1.8/15.8 GB
驱动程序版本:
驱动程序日期:
DirectX 版本:
物理位置:
14°C
晴朗
搜索
英
14:02
2024/3/13
2024/03/13 14:02:04

openvino.genai



intel Ai summit

openvino.genai

Benchmark tool

Input length – Output length

```
numactl -N 1 --membind=1 python benchmark.py -m /mnt/llm_irs/models_db24bab9_stateful/llama-2-13b-chat/pytorch/dldt/compressed_weights/OV_FP16-INT8_ASYM -d cpu -r ./test_report_db24bab9/ll
INFO:nncf:NNCF initialized successfully. Supported frameworks detected: torch, onnx, openvino
[ INFO ] ==SUCCESS FOUND==: use_case: text_gen, model_type: llama-2-13b-chat
[ INFO ] OV Config={'PERFORMANCE_HINT': 'LATENCY', 'CACHE_DIR': '', 'NUM_STREAMS': '1'}
[ INFO ] OPENVINO_TORCH_BACKEND_DEVICE=CPU
[ INFO ] Model path=/mnt/llm_irs/models_db24bab9_stateful/llama-2-13b-chat/pytorch/dldt/compressed_weights/OV_FP16-INT8_ASYM, openvino runtime version: 2024.0.0-14166-db24bab90ae-pr_22392
Compiling the model to CPU ...
[ INFO ] From pretrained time: 5.70s
[ INFO ] Read prompts from /home/build/jenkins_home/workspace/test-llm-model-imp/llm_prompts/32_1024/llama-2-13b-chat.jsonl
[ INFO ] Numbeams: 1, benchmarking iter nums(exclude warm-up): 3, prompt nums: 2
[ INFO ] [warm-up] Input text: Once upon a time, there existed a little girl who liked to have adventures. She wanted to go to places and meet new people, and have fun
[ INFO ] [warm-up] Input token size: 32, Output size: 128, Infer count: 128, Tokenization Time: 4.60ms, Detokenization Time: 0.40ms, Generation Time: 9.91s, Latency: 77.42 ms/token
[ INFO ] [warm-up] First token latency: 649.84 ms/token, other tokens latency: 72.89 ms/token, len of tokens: 128
[ INFO ] [warm-up] First infer latency: 649.01 ms/infer, other infers latency: 72.43 ms/infer, inference count: 128
[ INFO ] [warm-up] Max rss memory cost: 26407.07MBytes, max shared memory cost: 12568.50MBytes
[ INFO ] [warm-up] Result MD5:['c2ba8223fe81b826cf3e69f0b9af51c3']
[ INFO ] [warm-up] Generated: <s> Once upon a time, there existed a little girl who liked to have adventures. She wanted to go to places and meet new people, and have fun. She was always a
```

First Token Latency (ms)

Other Token Latency (ms)

Memory usage (GB)

intel Ai summit

openvino.genai

Benchmark tool

1. Convert HF models

CLI

```
python convert.py --model_id meta-llama/Llama-2-7b-chat-hf --output_dir models/llama-2-7b-chat
```

2. Run Benchmarking

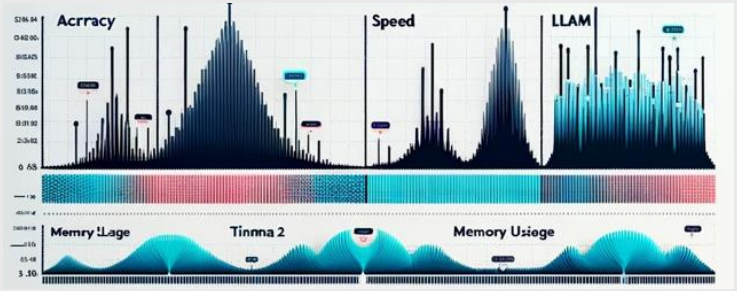
CLI

```
python benchmark.py -m models/llama-2-7b-chat/pytorch/dldt/FP32 -p "What is openvino?" -n 2
```

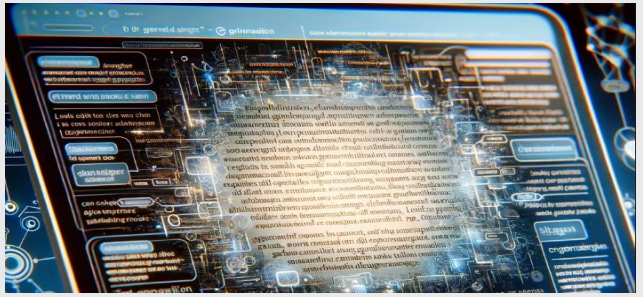
OpenVINO™ GenAI Pipeline Repo

OpenVINO Native C&C++ pipeline for Gen AI & LLM

github.com/openvinotoolkit/openvino.genai



Benchmarking for LLMs



Text generation C++ samples

Stable Diffusion (with LoRA) C++ image generation pipeline

LCM (with LoRA) C++ image generation pipeline

intel AI summit

OpenVINO™ 2024.0

Making it easier to deploy and accelerate Gen AI & LLMs

Improvement for LLMs

Simpler Workflow

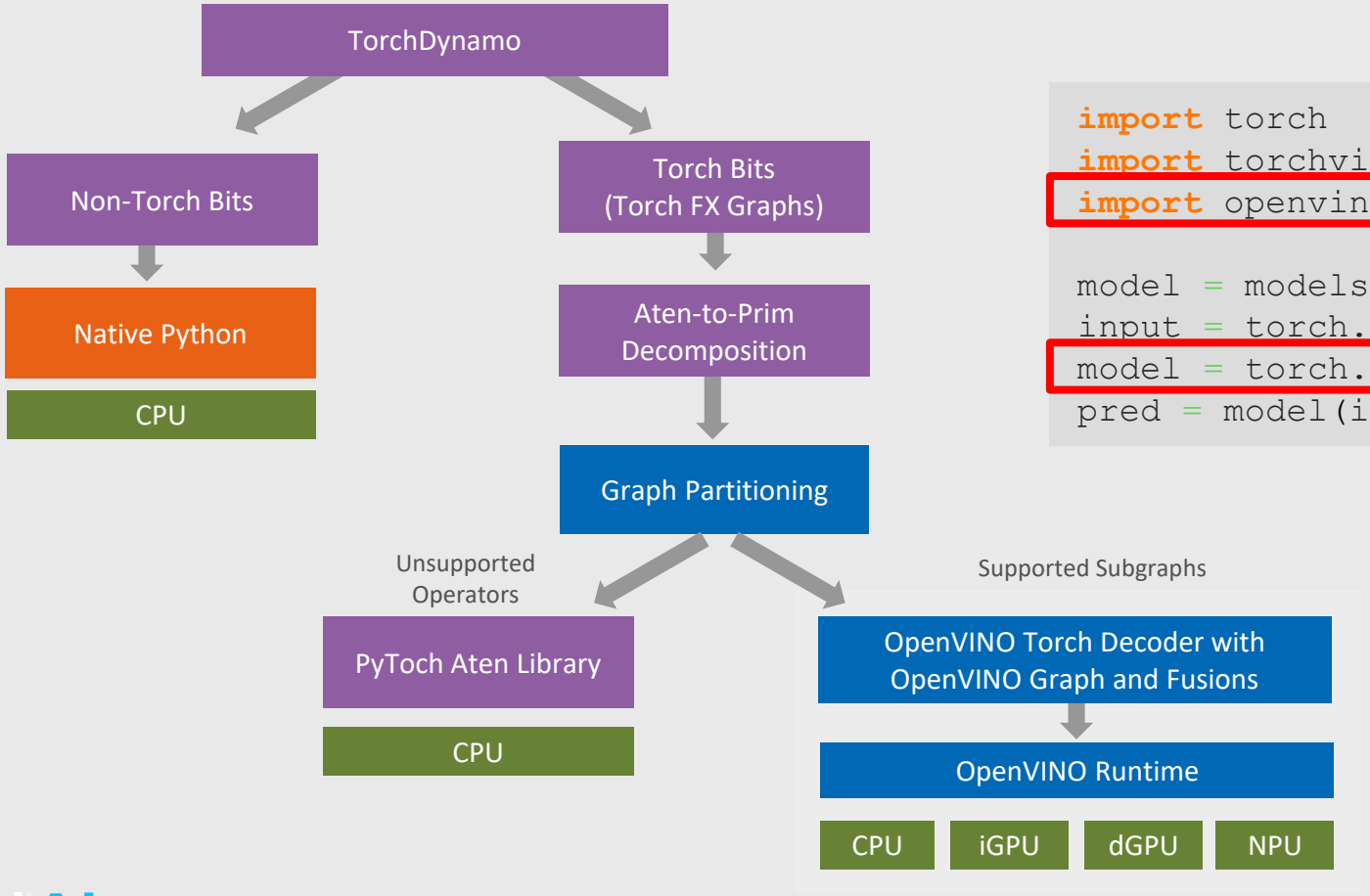
Enhancements for existing platforms

More Deep Learning Models

intel Ai summit

OpenVINO™ as Backend

PyTorch 2.0



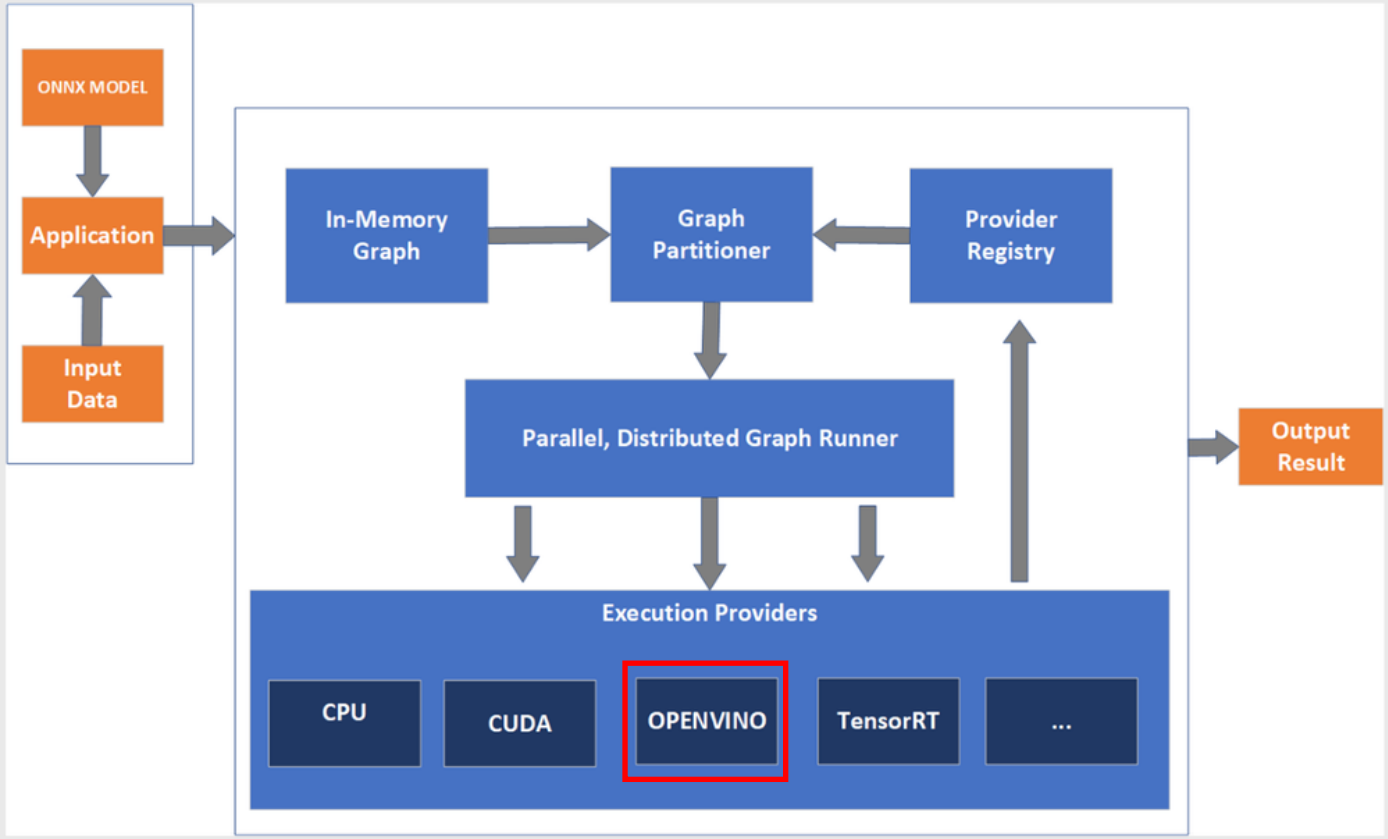

```python
import torch
import torchvision.models as models
import openvino.torch

model = models.resnet50(pretrained=True)
input = torch.rand((1,3,224,224))
model = torch.compile(model, backend="openvino")
pred = model(input)
```

Only 2 lines of codes

intel Ai summit

OpenVINO™ Execution Provider for ONNX



Installation

```
Linux
 pip install onnxruntime-openvino
Windows
 pip install onnxruntime-openvino
openvino
```

intel AI summit

JavaScript API

```
const { addon: ov } = require('openvino-node');
const { cv } = require('opencv-wasm');

const core = new ov.Core();
const model = await core.readModel('model.xml');
const compiledModel = await core.compileModel(model, 'CPU');

const imgData = await getImageData(imagePath);
const image = cv.matFromImageData(imgData);
const tensorData = new Float32Array(image.data);
const shape = [1, image.rows, image.cols, 3];
const inputTensor = new ov.Tensor(ov.element.f32, shape, tensorData);

const inferRequest = compiledModel.createInferRequest();
inferRequest.setInputTensor(inputTensor);
inferRequest.infer();

const outputLayer = compiledModel.outputs[0];
const resultInfer = inferRequest.getTensor(outputLayer);
const predictions = Array.from(resultInfer.data)
```

intel Ai summit

OpenVINO™ 2024.0

Making it easier to deploy and accelerate Gen AI & LLMs

Improvement for LLMs

Simpler Workflow

Enhancements for existing platforms

More Deep Learning Models

intel Ai summit

402-pose-est... (2) - JupyterLab

localhost:8888/lab/tree/402-pose-estimation-webcam/402-pose-estimation.ipynb

File Edit View Run Kernel Tabs Settings Help

254-rag-chatbot.ipynb | 402-pose-estimation.ipynt

Code | Python 3 (ipykernel)

Run the pose estimation:

[ ]:
```
USE_WEBCAM = True
cam_id = 0
video_file = "https://github.com/intel-iot-devkit/sample-videos/blob/master/store-aisle-detection.mp4?raw=true"
source = cam_id if USE_WEBCAM else video_file

additional_options = {"skip_first_frames": 500} if not USE_WEBCAM else {}
run_pose_estimation(source=source, flip=isinstance(source, int), use_popup=False, **additional_options)
```

[ ]:

Simple | 0 | 2 | Python 3 (ipykernel) | Idle | Mode: Edit | Ln 1, Col 10 | 402-pose-estimation.ipynb | 1

任务管理器

性能 | 运行新任务

CPU
5% 3.03 GHz

内存
11.7/31.6 GB (37%)

磁盘 0 (C: D:)
SSD
0%

以太网
以太网
发送: 0 接收: 0 Kbps

GPU 0
Intel(R) Arc(TM) Grap..
9%

NPU 0
Intel(R) AI Boost
0%

NPU
Intel(R) AI Boost

Compute 0% | Copy 0%

共享内存利用率 15.8 GB

利用率
0%

共享内存
0.1/15.8 GB

总内存
0.1/15.8 GB

驱动程序版本: 31.0.100.1688
驱动程序日期: 2023/10/31
DirectX 版本: 12 (FL 1.0 : Compu...
物理位置: PCI 总线 0、设备 1...

8°C 大部晴朗

搜索

17:13
2024/1/26

OpenVINO™ 2024.0

Making it easier to deploy and accelerate Gen AI & LLMs

- Improvement for LLMs
- Simpler Workflow
- Enhancements for existing platforms
- **More Deep Learning Models**

intel Ai summit

OpenVINO Notebooks

100+

Demos

LLMs, GenAI, Stable Diffusion, Whisper, GPT, YOLOv5/v8, CLIP,
Object Detection and Segmentation, Image Classification,
Human Pose Estimation, and much more!

intel Ai summit

OpenVINO Notebooks

Depth Estimation

intel Ai summit

OpenVINO Notebooks

PhotoMaker

'sci-fi, closeup portrait photo of a man img in Iron man suit, face'

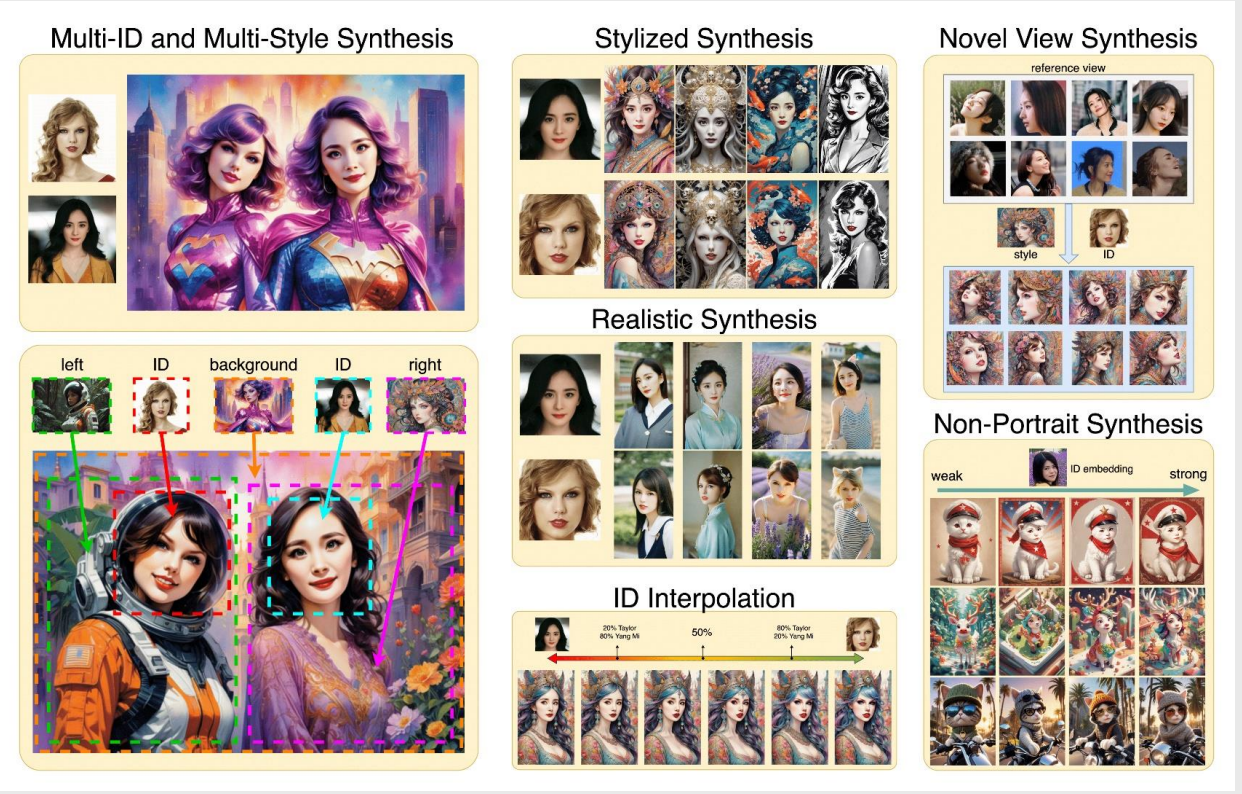
Original image

Output image

intel Ai summit

OpenVINO Notebooks

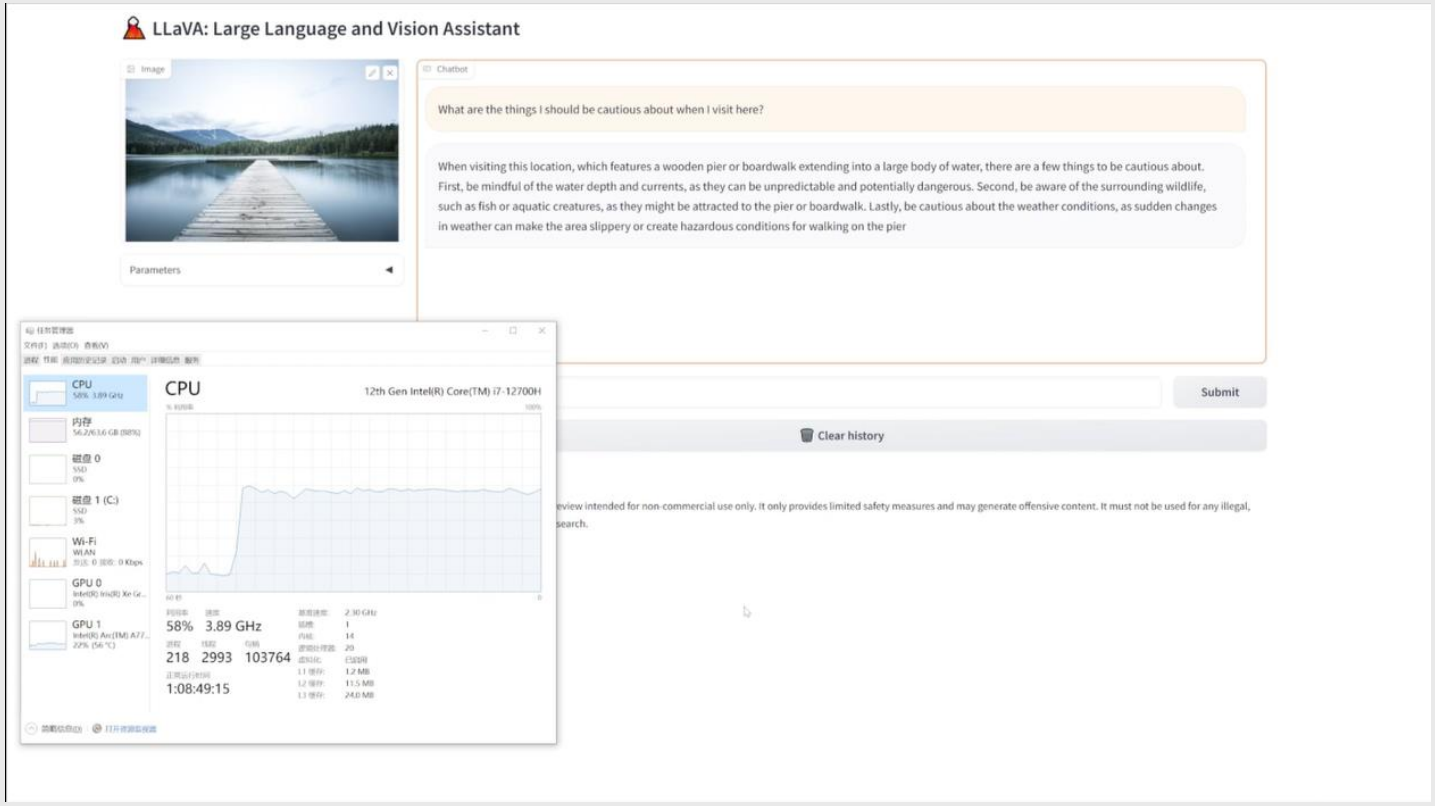
InstantID



intel Ai summit

OpenVINO Notebooks

LLaVA
(Large Language and Vision Assistant)

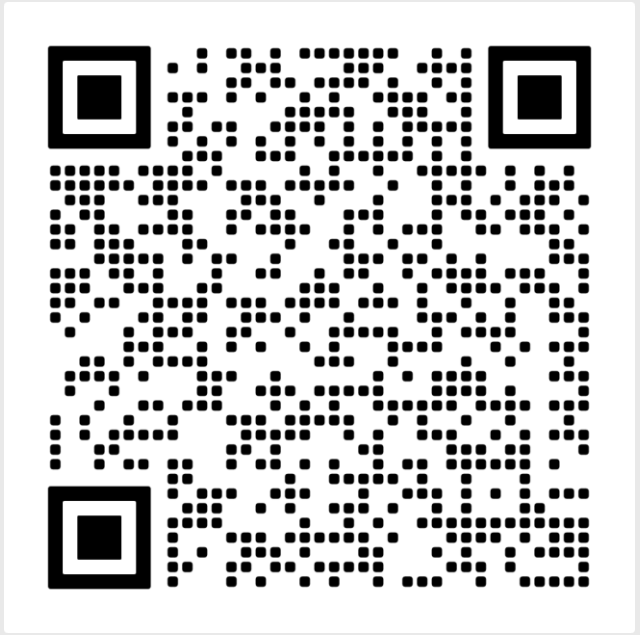


intel Ai summit

OpenVINO Notebooks

And Music Generation, Text to Speech (Bark), Speech to Text (Whisper), Diarization...

Installation



ubuntu

Red Hat

docker

ANACONDA

CONDA-FORGE

CONAN
C/C++ Package Manager

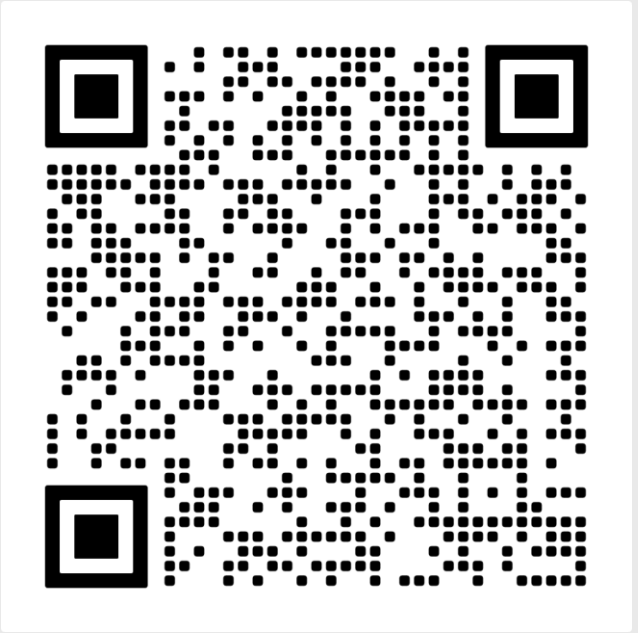
npm

www.openvino.ai

intel Ai summit

Installation

```
pip install openvino
```



ubuntu

Red Hat

docker

ANACONDA

CONDA-FORGE

CONAN
C/C++ Package Manager

npm

www.openvino.ai

intel Ai summit

Installation

Install OpenVINO™ 2024.0

Version

- **2024.0** Recommended
- Nightly Build
- 2023.3 LTS
- 2022.3.1 LTS Includes NCS2/HDDL support

Operating System

- **Windows**
- macOS
- Linux

Distribution

- OpenVINO Archives Includes NPU plugin
- **PIP** Includes NPU plugin Python API only
- GitHub Source
- Gitee Source
- Docker
- Conda
- vcpkg Source
- Conan
- npm JavaScript API only

Install

```
# Step 1: Create virtual environment
python -m venv openvino_env
```

```
# Step 2: Activate virtual environment
openvino_env\Scripts\activate
```

```
# Step 3: Upgrade pip to latest version
python -m pip install --upgrade pip
```

```
# Step 4: Download and install the package
pip install openvino==2024.0.0
```

Installation Instructions Previous Releases

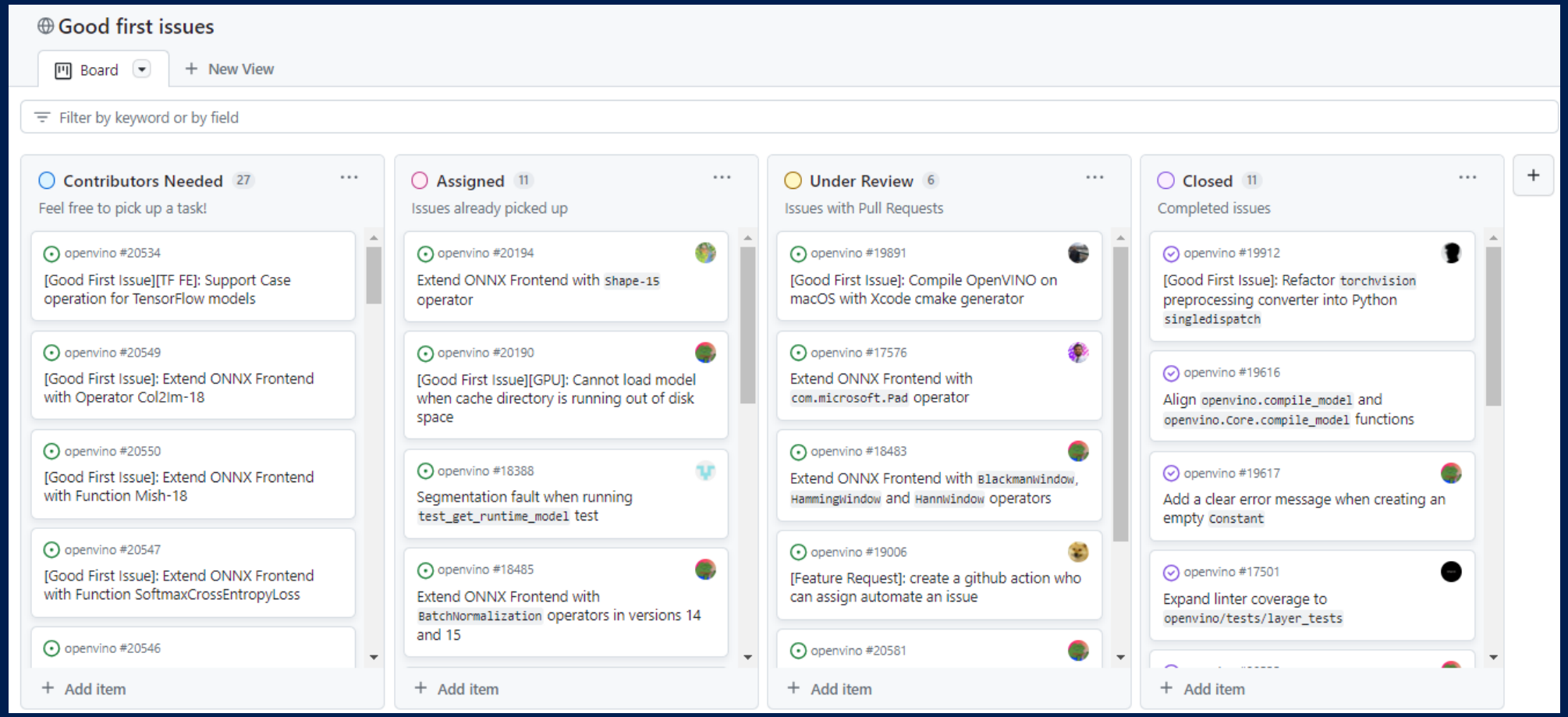
Advanced Optimization tool available separately: Learn about NNCF

Try It Yourself
openvino.ai
OpenVINO
intel

Connect With Us

intel Ai summit

Contribute to OpenVINO™ Toolkit



How to start: https://medium.com/openvino-toolkit/how-to-contribute-to-an-ai-open-source-project-c741f48e009e

intel Ai summit

Notices and Disclaimers

For notices, disclaimers, and details about performance claims, visit www.intel.com/PerformanceIndex or scan the QR code:

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Thank You!